DOI：10.7652/xjtuxb201604001

采用用户名相似度传播模型的线上用户身份属性关联方法

刘兆丽[1]，秦涛[1]，管晓宏[1,2]，赵丹[1]，杨涛[3]
（1.西安交通大学智能网络与网络安全教育部重点实验室，710049，西安；2.清华大学智能与网络化系统研究中心，100084，北京；3.西北工业大学自动化学院，710072，西安）

摘要：针对用户跨线上行为复杂多样难以融合监控的问题，提出了基于用户名相似度传播模型的线上用户身份属性关联方法。结合中文社交网络中用户名的特征，将用户名中的中英文字符进行分离，并采用贪婪算法分别求取不同用户名之间的中英文字符串的最大公共子串，以此实现含中英文字符的用户名相似度的计算；结合用户线上的好友结构网络，仅利用一阶邻居的用户名相似度求解用户对的匹配度，由此不但实现了用户名相似度沿网络结构的快速传播，也大幅度地降低了匹配算法的计算复杂度。结合所收集的新浪微博和人人网中用户身份属性数据的实验结果表明：新提出的字符串匹配算法将用户名匹配准确率提升了近30%，传播模型也大幅度地减少了用户名匹配的计算量，分析结果不但可以实现用户跨线上应用行为的关联融合，也对网络舆论控制和行为监管具有重要的参考价值。

关键词：线上应用；属性关联分析；用户名相似；特征传播
中图分类号：TP393 **文献标志码**：A **文章编号**：0253-987X(2016)04-0001-06

A Correlation Method of Online User Identity Attributes Based on a Propagation Model of Username Similarities

LIU Zhaoli[1], QIN Tao[1], GUAN Xiaohong[1,2], ZHAO Dan[1], YANG Tao[3]
(1. MoE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China;
2. Center for Intelligent and Networked Systems, Tsinghua University, Beijing 100084, China;
3. Department of Automation, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: A user identity attribute correlation method is proposed to focus on the problem that behaviors of online users are hard for fusion and supervision among multi-online applications due to their complexity and variation. The method is based on a propagation model of username similarities. The English and Chinese characters in usernames are separated by considering the characteristics of username in the Chinese social networks. A greedy algorithm is used to extract the longest common sequence between English and Chinese characters respectively for different usernames, and then username similarities are calculated. User's online connection structure and the username similarities of their first-order neighbors are used to decide the matching degree of the selected user pairs. Hence, not only the username similarity is propagated quickly among the connection networks, but also the complexity of matching calculation is greatly reduced. Experimental results based on the datasets collected from Sina Microblog and Renren networks show that the proposed algorithm improves the matching accuracy of usernames by about 30%,

收稿日期：2015-11-30。 作者简介：刘兆丽(1985—)，女，博士生；秦涛(通信作者)，男，副教授。 基金项目：国家自然科学基金资助项目(61221063,61403301,61402373)；中国博士后科学基金资助项目(2014M562417)。

and the propagation model greatly reduces the calculation complexity of username similarities. The analysis results achieve the goal of user's behavior fusion among different online applications, and have a reference value for online network security management and user's online behavior supervision.

Keywords: online application; attribute correlation analysis; identity similarity; characteristic propagation

多种多样的线上应用极大地丰富了人们的生活,人们通过这些线上应用获取所需的信息、分享个人观点、发表个人言论、共享生活心得。大量的线上应用在丰富了人们日常生活的同时,也给谣言、暴恐等信息的快速传播提供了温床。越来越多的恐怖和极端组织利用网络来传播他们的信仰言论,给社会安全带来了严重的威胁。用户在线上应用中发表各种言论所使用的 ID 是虚拟的身份属性,对这些虚拟的身份属性进行关联融合分析,并实现虚拟身份到物理人的映射,可为线上应用中舆论的安全监管提供丰富的信息。

自网络被广泛使用以来,针对网络用户行为分析的研究得到大量的关注。文献[1]研究了论坛中的线上用户多账号的关联,提出了 4 种匹配方法:基于字符串的匹配、基于书写习惯的匹配、基于发帖时间的匹配和基于社交网络结构的匹配。文献[2]提出采用语义分析的身份属性关联方法,不利用字符串的相似性实现身份属性的关联。文献[3]提出了从 Web 页面中提取用户的身份属性并将之应用于罪犯的追踪。在前期的研究工作中,作者提出了基于用户行为特征的身份属性关联方法,利用身份属性和 IP 地址的组合信息实现身份属性的关联[4]。可见,目前大多数分析方法通常仅选择用户节点信息或者网络结构信息其中之一进行身份属性关联,由于当前的网络应用中用户的身份属性信息有数千万条,如果仅仅通过分析节点信息的相似度,计算复杂度将大幅度提高,而如果仅仅采用网络结构信息而不使用节点信息,那么身份属性匹配的准确度将大幅度降低。

针对上述问题,本文提出了基于用户名相似度传播模型的用户身份属性关联方法,首先结合中文社交网络中用户名的特征,将用户名中的中英文字符进行分离,分别求取不同用户名之间的中英文字符串的最大公有子串,借以实现用户名相似度的计算。随后结合用户线上的好友结构网络,仅利用一阶邻居的用户名相似度求解用户对的匹配度,由此不但实现了用户名相似度沿网络结构的快速传播,也大幅度地降低了匹配算法的计算复杂度。结合所收集的新浪微博和人人网中用户身份属性数据的实验结果表明,本文所提出的方法因综合利用用户节点和网络结构信息,在降低计算复杂度的同时大大提升了身份属性关联的效率。

1 用户名相似度传播模型

1.1 用户身份属性关联建模

在线社交网络中用户之间的关系可以采用图 $G(V,E)$来描述,其中 V 为用户的集合,E 为用户之间关系的集合。边的构建原则为用户之间存在好友关系,例如在微博网络中用户 $u,v\in V$,且用户 u 关注了用户 v,则存在一条从 u 指向 v 的有向边,即 $e_{uv}\in E$。目前线上应用类型很多,一个物理用户可能同时在多个不同的线上应用中拥有账户,也即是同一个物理人拥有多个线上身份属性。线上用户身份属性的关联可以定义为如何在两个线上社交网络 $G_1(V_1,E_1)$和 $G_2(V_2,E_2)$中,挖掘属于一个物理人的账号,既挖掘关联对(v_1,v_2),它们属于同一个物理人并且 $v_1\in V_1,v_2\in V_2$。

1.2 用户名相似度问题建模

用户名即用户账号,本文中称为用户的网络身份属性,它是由字母、汉字、数字以及一些特殊符号组成的字符串。在同一种网络应用中用户名作为用户的唯一身份标识,是不允许重复的。Kumar 的研究表明,50%的用户在不同的社交网络中使用相同用户名[5]。Bekkerman 的研究表明,线下物理人通常会在线上社会中使用相同或相近的用户名[6]。据此,可以使用用户名的相似性作为多网络用户关联的重要依据。

用户名的表现形式为字符串,因此多网络用户身份属性的关联问题可以转换为计算用户名之间的相似度问题。字符串相似度指的是寻找两个字符串的公共子串,利用公共子串的长度根据相应的公式来衡量两个字符串的相似程度[7]。对于待比较的两个字符串,分别用集合 S_1 和集合 S_2 表示。令 $S_1=\{s_{10},s_{12},s_{13},s_{14},\cdots,s_{1m}\}$,其中 $s_{1i}(i=0,1,\cdots,m)$依

次表示顺序构成字符串 S_1 的每一个字符；令 $S_2=\{s_{20},s_{22},s_{23},s_{24},\cdots,s_{2n}\}$，$s_{2j}(j=0,1,\cdots,n)$表示构成字符串 S_2 的每一个字符。$s_{1i}(i=0,1,\cdots,m)$与 $s_{2j}(j=0,1,\cdots,n)$之间所有的匹配对(s_{1i}^m,s_{2j}^m)构成的集合为 M，则式(1)定义了 S_1 与 S_2 之间的相似度 S。由定义可以看出，当 M 中元素越少，即匹配对越少，字符串 S_1 与 S_2 之间的相似度越小。当且仅当$|M|=|S_1|=|S_2|$时，S_1 与 S_2 的相似度为 1。

$$S(S_1,S_2)=[(|\{s_{1i}^m|(s_{1i}^m,s_{2j}^m)\in M\}|+|\{s_{2j}^m|(s_{1i}^m,s_{2j}^m)\in M\}|)/(|S_1|+|S_2|)]\times 100\% \tag{1}$$

1.3　用户名相似度的传播模型

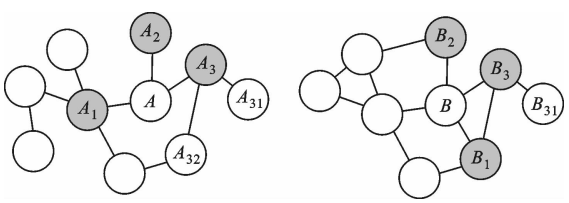
假设在所要匹配的两个线上应用中存在已经确定关联的用户身份属性，称这些属性为种子信息。种子信息是通过人工查询获得的同时存在于两个社交网络的账号对，利用种子信息可以增加算法的可信度。

在线上应用中用户有不同的好友，这些好友有可能是该用户在物理世界中的好友，也有可能是该用户在网络世界中的好友。根据用户的活动规律，我们推测用户在物理世界中的好友在不同的应用中很可能也应该是用户的好友。为此，我们得到同一物理用户在不同的应用中的好友会存在比较多的物理匹配对，并提出了结合网络结构的用户身份属性关联传播模型。传播模型采用基于网络结构的迭代算法，以种子集合作为初始已匹配对，从种子节点将其携带的信息扩散到其邻居节点，根据网络拓扑结构依次传播下去，最终得到两网络中相关联的用户。传播模型的计算步骤可简单描述为：

步骤 1　从种子集合中取出两个已匹配的节点，例如，如图 1 所示的 A 点和 B 点；

步骤 2　获取该对用户在各自网络中的邻居节点，例如图 1 中的 A_1、A_2、A_3 和 B_1、B_2、B_3，若种子集合中所有的匹配节点对都已经被计算，则计算结束；

步骤 3　针对步骤 2 获取的两个邻居节点集合，计算所有可能匹配对的相似度，在本例中需要计算相似性的节点对包括(A_1,B_1)、(A_1,B_2)、(A_1,B_3)、(A_2,B_1)、(A_2,B_2)、(A_2,B_3)、(A_3,B_1)、(A_3,B_2)以及(A_3,B_3)；



(a)网络 A　　(b)网络 B

图 1　结合网络结构的传播模型

步骤 4　在步骤 3 所计算的结果中选取满足一定条件的匹配对，同时将匹配对添至种子集合；

步骤 5　返回步骤 1。

2　线上行为信息收集与预处理

从 2013 年 8 月 15 日至 2013 年 9 月 14 日，我们先后采集了新浪微博中 1 808 600 名用户的相关信息以及他们的好友关系网络，从 2014 年 4 月 22 日至 2014 年 5 月 15 日共采集了人人网中 40 330 名用户的相关信息以及他们的好友关系网络信息。

2.1　用户噪音节点清除

为了提高后续关联分析的准确率，需要对现有数据进行去除领袖节点以及劣质用户。本文将新浪微博的用户分为 5 大类：名人、人气博主、官方微博、普通用户与劣质用户。名人是通过新浪实名认证、在各行各业里面具有一定知名度的人，包括明星(例如姚晨)、在较大型的公司企业里面担任一定职务的人(例如马云、雷军)等。此类人通常有较多的粉丝。人气博主是指发表微博主题明确、内容新鲜独到，吸引大量用户关注，大部分未提供个人真实信息的用户，通常包括两类用户，以个人为单位的网络写手(网络用语称为段子手，例如回忆专用小马甲、王思聪，行文幽默新颖)和以主题为单位而非个人的微博写手(例如全球热门收集、IT 程序猿)。官方微博是指各行各业由相关负责人进行主题与本单位相关微博的更新，发微博目的通常为品牌传播、产品营销、用户互动、危机预防与处理等。其中包括政府机关官方微博(例如江宁公安在线)、媒体官方微博(例如华商报)、企业官方微博(例如果壳网)等。普通用户是指使用网络进行信息发布以及信息获取的普通网络用户。劣质用户是指对微博文化有着不良传播和影响的用户，主要指广告用户和僵尸粉用户。

为了区分上述不同类型的用户，本文提出用户类型参数 α_{folk} 对用户进行分类，实现领袖节点、非个人用户以及广告或者僵尸粉用户的排除。α_{folk} 定义为用户的粉丝数与关注数的比值。在数据集中选取了名人、人气博主、广告与僵尸节点各 15 个，同时选取了普通用户节点 30 个，这些用户的粉丝数、关注数和 α_{folk} 的平均值的统计结果如表 1 所示。其中普通用户的粉丝数是 1 785.5，虽然这个数据不具有代表性，但是它可以有效地区分名人和普通用户。在

前期的研究工作中，我们对大量普通用户的测量分析发现，只有 10%左右的普通用户的粉丝数大于 1 000[8]，但是在本文中，为了综合考虑并实现普通用户和名人的差别，我们也选择了一些粉丝数比较多的普通用户作为样本数据，使得普通用户的粉丝数较多。

由表 1 的结果可以看出，名人节点等的用户类型参数值较大，而僵尸节点与广告节点的该值均小于 0.1，而普通用户的用户类型参数往往大于 0.3。因此，本文可使用用户参数类型来解释不同类型的用户：当 α_{folk} 在[0.3,50]区间时，用户为普通用户节点；当 α_{folk} 在(0,0.3)时，用户为僵尸节点或者广告节点；当 α_{folk} 在(50,n)时，用户为名人、人气博主等节点，其中 n 为微博粉丝上限数。通过上述方法选取普通用户节点，在绝大多数情况下排除了名人、人气博主与广告、僵尸粉用户，同时保留了活跃度高的用户以及小有名气的用户。

与此同时，我们利用人工标定的方法在新浪微博和人人网中标记了 27 对匹配节点，它们组成了种子节点集合。

表 1 不同类型用户的类型参数值

用户类型	粉丝数均值	关注数均值	α_{folk} 均值
领袖节点	14 720 670	502	29 324
僵尸节点	17.8	241	0.073 9
广告节点	88	1 015	0.086 7
普通用户	1 785.7	266.6	6.698 0

2.2 用户好友网络构建

本文定义的用户关系网为用户的好友关系，即在线下生活中相互认识的朋友关系或者在网络上成为朋友关系。在新浪微博中，用户 A 关注用户 B，即 A 是 B 的粉丝，并不意味用户 A 与 B 就是好友关系。当两个人同时关注了对方，即 A 关注 B 同时 B 也关注 A，则在大部分情况下互为朋友关系，因此用户是否相互关注可以用来判断用户是否存在好友关系。此外，社交关系模型的建立可以将新浪微博原有的有向用户关系转化为无向关系。对于人人网，只有当两用户均为对方的好友时，才会在网页上显示其互为好友关系，因此人人网中以用户为节点，以好友关系为边，所构成的网络图为无向图，无需对网络结构进行预处理。此外，人人网单个节点均为一个线下物理人，人人网中所存在的非个人账号，即公共主页，并不在好友列表中，因此无需进行排除。

3 用户名相似度度量方法

3.1 用户名相似度计算方法

用户名相似性计算可以转化为字符串相似性计算，而字符串相似性计算方法中最为常见的包括编辑距离算法(Levenshtein)、最长公共子序列(longest common subsequence，LCS)算法和贪心算法(greedy string tiling，GST)算法。其中编辑距离算法是由俄国科学家 Levenshtein 首先提出，故又称作 Levenshtein 距离，编辑距离是指两个字符串之间，由一个字符串转换成另一个字符串所需要的最少编辑次数。一次编辑的合法操作包括：将某字符替换为另一字符，插入某字符以及删除某字符[9]。LCS 算法也是计算字符串相似度的常用算法。该算法将两个给定的字符串分别删除零个或者多个字符，但不改变剩余字符串的顺序而得到的长度最长的相同子字符序列[10]。其中子序列是原字符串的子串且保持每个元素在原字符串中相对位置不变[11]。GST 算法是一种贪婪算法，最早由澳大利亚悉尼大学的 Michael Wise 设计。该算法对两个字符串进行贪婪式搜索以找出最大公有子串，需要对字符串进行多次搜索，每次找出当前字符串中未标记部分的最长公共子串，同时将已找出的最长公共子串标记为已使用，避免最大匹配重复使用[12]。

3.2 针对中文社交网络用户名特征的相似度算法

GST 算法会在以下情况中出现较大误差：当用户名中包含较多无关信息时，如特殊符号或无意义字符，算法计算的值较低；当用户名为繁体字时，难以使用原算法进行匹配，致使失配率增高；当设置最短匹配长度(MinMatchLength)取值较低时，会造成英文字符串的匹配较多，尤其当英文元字母较多，如“Ainne”与“Irene”，当 MinMatchLength 取值为 1 时相似度为 0.8，而当 MinMatchLength 取值为 3 时，相似度为 0。然而，用户名往往不会仅仅包含英文，多为中文或者中英文结合。为了进一步提高算法准确率，分析发现用户在社交网络中的用户名存在以下特征：①社交网络用户由于其存在部分的真实线下社交特点，有部分用户会将自己的线下真实姓名用于用户名之中，并多倾向于使用真实姓名作为用户名的子序列；②用户的用户名中，同时包含部分与真实姓名无关的字符；③用户中文名为姓氏与名字所构成，姓氏最少是一个字，用户名如果包含真实姓名，则其中最小可信单位长度为 1，英文作为人名时，最小长度为 3，则可信最小单位长度为 3。基

于以上分析，本文对字符串相似度算法 GST 进行改进。改进后算法命名为 Chinese greedy string tiling，简写为 CGST，具体步骤如下。

步骤 1　排除对用户名匹配时贡献小的字符，即对两个字符串去除数字及特殊符号字符。

步骤 2　转换字符串中繁体字为简体字。

步骤 3　如果当一个字符串包含另一个字符串，则相似度可设置为一个较高的固定值，如 1.0；若不包含，则转入步骤 4。

步骤 4　分别计算两字符串的中文字符串最长公共字串集合与英文字符串最长公共字串集合，计算中文字符串时，MinMatchLength 取值为 1；计算英文字符串时，MinMatchLength 取值为 3。

步骤 5　计算相似度。在获取最长公共字串集合的算法中，每次在向当前循环中添加新的字符时，需要判断该字符是否已经匹配过，以排除重复匹配情况。

4　实验结果分析

4.1　相似度方法比较

将字符串相似度应用于用户名比较时，评判算法是否合适的衡量依据主要为计算结果是否准确，即能否使同一个物理用户在两个应用中的用户名计算得出的相似度值高。因此，下文中主要通过应用算法到具体场景来比较 3 个算法的性能。在数据集中，作者标定了 27 位同时使用新浪微博与人人网的物理用户，作为种子。在此选取其中 3 位用户将其用户名列出，见表 2。该表中的用户将用于分析不同算法的准确性。

表 2　部分种子用户跨应用用户名

编号	用户名	
	新浪	人人网
1	冯嘉豪 Rock	冯嘉豪 Rock
2	小雯雯-滕雯	滕雯
3	田若静-不美不开心	田若静

采用第 3 节中的不同算法分别对这 27 个用户的两个用户名进行相似度计算，图 2 展示了不同算法计算的相似度分布结果。图中横坐标表示相似度，纵坐标表示用户数，每一个点表示使用不同算法大于该相似度的用户数。对于确定已匹配的用户名对，相似度越高的用户越多表示算法准确率越高。可以看出：编辑距离算法所得到的用户名相似度多数小于 0.4，占总用户数的 61.5%；LCS 算法较编辑距离算法来说结果略好一些，69.2%的用户计算所得相似度小于 0.5；GST 算法相似度为 0.4 以上的用户较多，占总用户数的 77%。通过计算得到，3 个算法计算所得的平均相似度值分别为 0.42、0.44 和 0.58，由此可看出 GST 算法在计算用户名相似度时有较好的表现。对于表 2 中编号为 2 的用户，新浪微博用户名为“小雯雯-滕雯”，人人网用户名为“滕雯”，通过编辑距离算法、LCS 算法和 GST 算法计算的相似度差异较大，分别为 0.17、0.36 和 0.89。编辑距离算法计算得到的值很低主要因为两个字符串长度差异大，导致即使包含相同的字符串时也会使差异步数较大，从而计算结果不理想。LCS 的结果主要由于当存在特殊符号隔开了有意义的字符时，会导致计算结果下降。GST 算法反复计算了两字符串的公共字串，致使这对字符串有很多匹配对，因而计算结果较为理想。从算法所匹配的公共序列来看，编辑距离算法与 LCS 算法均属于有序的相似度算法，而 GST 算法属于无序的算法，即对于 GST 来说，“王小红”与“小红王”是完全匹配的，而对于前两种算法无法完全匹配，这也是其算法准确率较低的原因。

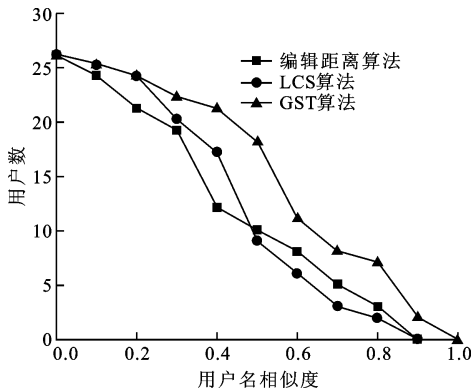


图 2　编辑距离、LCS 及 GST 算法的计算结果比较

4.2　CGST 算法性能分析

对种子集合中的 27 对用户名采用 CGST 算法和 GST 算法的计算结果对比如图 3 所示，有 80.8%的用户名对的相似度大于等于 0.9，而 GST 算法在此区间仅有 7.7%的用户。GST 算法用户名平均相似度值为 0.58，而 CGST 算法把结果提高为 0.89。编号为 3 的用户的新浪微博名与人人网用户名分别为“田若静-不美不开心”与“田若静”，采用 GST 算法得到相似度值为 0.5，而采用 CGST 则为 1.0。例如“何鹏程”和“郝程程程”这个并非属于同一个物理人的用户名对，相似度计算的结果由 0.86 降低为 0.29。CGST 算法大大提高了用户名的区分度。本文提出的 CGST 算法与 GST 算法相比，改进主要

在于每一轮计算中会保存已经添加的匹配对，每次新检测出的匹配对需要查询是否已存在该匹配对。如果采用哈希表等数据结构进行存储，每次查询的时间复杂度降为 $O(1)$，因此改进后的 CGST 算法并没有增加时间复杂度。当两个字符串完全不相同时，计算的时间复杂度为 $O(n^2)$；当每次循环中均只能找到一个最大匹配，将致使算法 3 个循环均要被执行，则时间复杂度为 $O(n^3)$。但是，鉴于用户名匹配仅仅在用户的邻居节点集合进行，而 90% 以上的用户其好友个数均小于 1 000[8]，因此本算法的计算复杂度并不高。

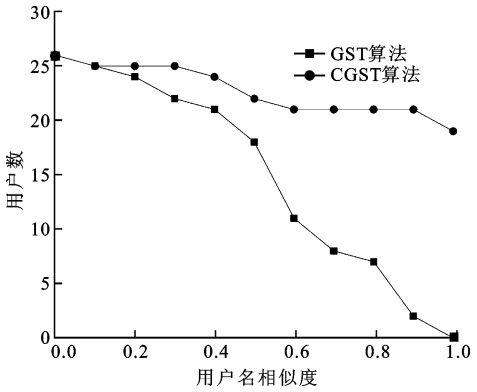


图 3 GST 与 CGST 算法比较

4.3 传播算法计算复杂度分析

传播算法每一轮选取一对已匹配对，在两个网络中遍历匹配对的所有好友，该过程需要 $O(n^2)$ 的时耗。基于 CGST 的用户名相似度计算方法的复杂度为 $O(n^3)$，其中 n 为用户名长度，这里的 n 与数量级上百的遍历好友数目相比较小，可忽略不计。传播算法的总轮数由新匹配数来决定，而新匹配数与网络规模有关，当总轮数与好友数的数量级相近时，基于用户名的传播关联方法的时间复杂度为 $O(n^3)$，当总轮数远远大于用户好友数时，则可忽略遍历好友的时间，即时间复杂度为 $O(n)$，由此可见本文所提出的算法计算复杂度较低。此外，本文所提出的基于用户名的传播关联方法，将用户名相似性的计算与网络拓扑结构相结合，摒弃了以往在两个网络中直接进行用户名相似度计算的方法，从而避免了当数据量多时用户名重复率高，导致用户名计算结果相似或者区分率小、准确率低以及计算复杂度高的问题。

4.4 算法匹配准确度分析

本文使用新浪微博 37 608 名用户和人人网 40 330名用户及其好友关系生成的网络，以 27 名用户作为种子节点进行节点匹配计算，最终得到 276 名物理用户的两网账号匹配对，见表 3。经过对其中 150 对关联进行人为查询，发现当阈值选取 0.75 时，得到的匹配准确率高达 96.77%。

首先，这是因为我们在采集数据时，新浪微博爬虫的起始点和人人网络爬虫的起始点不是物理匹配对，造成了两个网络中总体的覆盖率不是很大，如果以种子节点为起始节点，匹配率将会大幅度提升。此外，匹配结果与节点和种子节点所处的网络位置有关，当节点处于网络边缘，则缺失节点结构信息，会导致计算结果不佳。最后，当网络较小时，处于网络边缘的节点数占网络规模比值大，也会造成计算结果不佳。如果增加网络规模或者种子节点的数量，那么匹配率也会得到提升。

表 3 关联方法关联结果信息

应用平台	节点数	种子数	种子占比/%	匹配数	匹配率/%
新浪	37 608	27	0.071	276	0.73
人人网	40 330	27	0.066	276	0.68

5 结 论

线上应用在丰富了用户生活的同时，也给网络舆论安全管理带来一定的困难，本文围绕用户跨线上应用身份属性关联问题开展了研究。首先，本文对所收集到的新浪微博信息和人人网信息进行了预处理，清除了新浪微博中的僵尸粉等账号。再者，根据中文社交网络中用户名的特点，提出了基于 CGST 算法的用户名相似度计算方法。最后，为了充分利用用户节点的信息和网络结构方面的特征信息，本文提出了基于用户名相似度传播模型的身份属性关联方法，在大幅度降低计算复杂度的同时，也有效地提升了身份属性信息关联的准确性。结合所收集到的用户身份属性信息的实验结果表明，本文所提出的方法在计算复杂度和匹配准确度方面都具有较高的性能。在下一步工作中，将综合考虑更多的用户身份属性信息，进一步提升匹配的准确度。

参考文献：

[1] JOHANSSON F, KAATI L, SHRESTHA A. Detecting multiple aliases in social media [C]// Proceedings of the International Conference on Advances in Social Networks Analysis and Mining. Piscataway, NJ, USA: IEEE Communication Society, 2013: 1004-1011.

（下转第 27 页）